


Performance of non-binary athletes in mass-participation running events

John Armstrong ¹, Alice Sullivan,² George M Perry³

To cite: Armstrong J, Sullivan A, Perry GM. Performance of non-binary athletes in mass-participation running events. *BMJ Open Sport & Exercise Medicine* 2023;**9**:e001662. doi:10.1136/bmjsem-2023-001662

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjsem-2023-001662>).

Accepted 4 December 2023

ABSTRACT

Objectives To test the hypothesis that, controlling for age, natal-sex differences in running performance are lower among non-binary athletes than in the rest of the population. To test the hypothesis that natal-male non-binary athletes outperform natal-female non-binary athletes.

Methods A secondary analysis of 166 race times achieved by non-binary athletes within a data set of 85 173 race times derived from races with a non-binary category in the New York Road Runners database. The natal sex of non-binary athletes was modelled probabilistically using US Social Security Administration data when it could not be derived from previous races. Race times were used as the outcome variable in linear models with explanatory variables derived from natal sex, gender identity, age and the event being raced. Statistical significance was estimated using Monte Carlo methods as the model was not Gaussian.

Results There was no evidence that controlling for age, natal-sex differences in running performance are lower among non-binary athletes. Natal-male non-binary athletes outperform natal-female non-binary athletes at a confidence level of $p=0.1\%$.

Conclusions Both natal sex and gender identity may be useful explanatory variables for the performance of athletes in mass-participation races. It is, therefore, valuable to include both variables in data collection.

INTRODUCTION

Policy debates regarding the relative importance of sex and gender identity have taken place across a wide range of domains, including sports. The literature addresses questions of fairness and participation within sports, particularly relating to the female sports category.^{1–5} In contrast, the current paper seeks to empirically test the respective predictive power of gender identity and natal sex on running performance using the non-binary category. In recent years, several running events have introduced a non-binary category alongside the male and female categories on the grounds that people who identify as non-binary ‘cannot compete authentically within the existing system’.⁶ However, this paper is not intended to study the merits of a non-binary category. Our

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Outside of purely biological outcomes and criminology, little empirical work has been done to test the theory that gender identity is more important than natal sex as a cause of gender disparities in outcomes.

WHAT THIS STUDY ADDS

⇒ In mass-participation running, identifying as non-binary does not reduce gender disparities. This provides evidence against the theory that an individual’s gender-identity plays a significant role in these disparities in addition to their natal sex.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The study highlights the importance of collecting data on both natal sex and gender identity.

interest in the category arises because it gives us a data set that can be used to test gender-identity theory.

It is contested whether sex is ‘binary’,^{7 8} a ‘multidimensional biological construct’,⁹ or whether ‘the ‘naming’ of sex is an act of domination and compulsion, an institutionalised performative that both creates and legislates social reality by requiring the discursive/perceptual construction of bodies in accord with principles of sexual difference’.¹⁰ For this paper, it is not necessary to take sides in this debate. We will use ‘sex’ to refer to sex registered at birth, and we use the terms natal male and natal female to refer to sex registered at birth.

According to gender-identity theory, everybody has an innate gender identity, defined by UK LGBTQ+ (lesbian, gay, bi, trans, queer, questioning and ace) Charity Stonewall as ‘A person’s innate sense of their own gender, whether male, female or something else ... which may or may not correspond to the sex assigned at birth.’¹¹ Advocates of gender-identity theory argue that gender identity is typically more important than sex in determining outcomes, which are socially influenced, an idea sometimes expressed as ‘trans women are women’.¹² For example, according to Safer,¹³ ‘for general



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY. Published by BMJ.

¹Mathematics, King’s College London, London, UK

²University College London, London, UK

³Independent Researcher, Houston, Texas, USA

Correspondence to

Dr John Armstrong;
john.armstrong@kcl.ac.uk



categorisation by sex, 'brain sex' or gender identity would be the default characteristic'. Such viewpoints have been widely adopted, leading to some statistical bodies advising that data on sex should not be collected.^{14–18}

Yet the question of the respective importance of sex and gender identity on relevant outcomes and the interaction between the two is an empirical one that cannot be answered without data on both gender identity and sex. The existing literature is limited and, in both sports and criminology, typically relates to the effects of medical transition rather than identity per se.^{19–21}

Running counter to gender-identity theory is the gender-critical belief 'that biological sex is real, important, immutable and not to be conflated with gender identity'²² and that sex often has important effects that cannot be explained purely by gender identity. We acknowledge that not everyone who holds these views would choose to label themselves as gender critical.

We will consider sporting performance as a test case for these theories. In mass-participation events, social factors, such as levels of training, personal motivation, personal expectation and natural competitiveness, are likely to impact sporting performance significantly. Even for elite athletes who do not lack motivation and competitiveness, there are likely to be gendered differences caused by issues such as family commitments, the availability of sponsorship and levels of access to coaching and individualised sports science. Such issues have been studied extensively in the literature on gender and sport.^{23 24} The potential importance of social determinants in sports is also highlighted in.⁹ Suppose gender identity influences how one performs social roles.¹⁰ Since the sports literature indicates that gendered roles influence sports performance, in that case, one should expect that gender identity may be associated with sports performance.

We will study the application of gender-identity theory to road-racing data gathered by New York Road Runners (NYRR), whose races in recent years have included three categories: male, female and non-binary. Their non-binary category is a simple matter of self-identification as either identifying as non-binary or not, though non-binary identities are varied and complex.^{25 26} One definition of non-binary is 'an adjective describing a person who does not identify exclusively as a man or a woman. Non-binary people may identify as being both a man and a woman, somewhere in between, or as falling completely outside these categories. While many also identify as transgender, not all non-binary people do. Non-binary can also be used as an umbrella term encompassing identities such as agender, bigender, genderqueer or gender-fluid.'²⁷ The qualitative literature on non-binary identities and sports suggests that people who identify as non-binary face barriers to participation but typically does not interrogate how these experiences vary by sex.^{28–31}

METHODS

We will test hypotheses derived from gender-identity theory and gender-critical theory by examining how

similar the performance of athletes in the non-binary category is to that of other athletes of the same sex.

Our first hypothesis derives from gender-identity theory.

Hypothesis 1

Controlling for age, sex differences in non-binary athletes' race times will be smaller than the sex differences in race times observed for other athletes.

Our second hypothesis derives from the gender-critical view.

Hypothesis 2

Controlling for age, natal-female non-binary athletes will tend to have slower race times than natal-male non-binary athletes.

Note that hypothesis 1 is derived from, but not equivalent to, gender-identity theory. Similarly, hypothesis 2 is derived from, but not equivalent to, gender-critical theory. Thus, these hypotheses are natural tests to perform from the point of view of falsifying each of the theories. Note also that these hypotheses are not mutually exclusive: it is possible that both sex and gender-identity have significant effects.

Unfortunately, the sex of athletes who identify as non-binary is not recorded in our data set. To overcome this obstacle, we use the novel technique of modelling the likely sex of athletes based on their given names. We do not claim that one can perfectly predict sex based on an athlete's given name. Instead, we develop a probability model for sex. Our technique is similar to the accepted statistical practice of imputing missing values in a data set.³² We will show through cross-validation that our probability model can be used to predict the natal sex of non-binary athletes. In online supplemental appendix C, we repeat our analysis using a probability model that includes additional uncertainty from causes such as athletes changing their names. This increases the estimate for the size of sex differences among non-binary athletes and, if one assumes sufficient uncertainty would suggest a statistically significant increase in sex differentials, the opposite effect to that predicted by gender-identity theory. This should be viewed simply as showing that our model is not skewed in favour of gender-critical theory.

We also performed exploratory analyses of other potential associations with gender identity, discussed below.

Data

We used race data from the NYRR database,³³ selecting races that featured non-binary athletes giving 21 races listed online in online supplemental appendix A. This data set was selected as the largest available consistently formatted data on non-binary athletes. Before embarking on the study, we performed a power analysis that indicated the data set would be sufficient to detect if the times of natal-male and natal-female athletes were equal in the non-binary category (based on NYRR marathon data, we estimated that the difference in the mean race times was

24.6 min with an SD of 58 min, we estimated that a total sample size of 138 was required to achieve 80% power at 95% CI for a one-tailed t-test). Where possible, the sexes of non-binary athletes with a given name were identified using the Athlinks website.³⁴ The frequencies of different baby names in each year in the USA were obtained from the US Social Security Administration baby name database.³⁵

Creation of cross-race data set

Because we were using data from multiple races, the same athlete had often competed in multiple races. We assumed two athletes of different races with the same name were the same athlete.

We wished to create a data set, which contained only one record for each athlete. We decided to base our decision on which result to choose for a given athlete by choosing the race most closely correlated to the marathon once outliers had been discarded. Our assumption that athletes with the same name were the same athlete may have led us to unnecessarily discard some records reducing the sample size and statistical power slightly.

In more detail, for each race, we identified the athletes who had also run the New York Marathon. We then fit a linear model with no intercept, which allowed us to estimate an athlete's marathon time based on their race time. We then fit a second linear model to the 98% of times having the lowest values for the ratio of the residual to their time. These second linear models allowed us to predict any runner's marathon time based on their time in the race; we will just call this the predicted marathon time. See figure 1 A for an example. We will call the correlation coefficient of this time the marathon correlation of the race.

Table 1 Summary of the number of records used in the creation of the cross-race data set

Total no of race results in full data set	186 782
No of outliers discarded	159
No of unique rows in cross-race data set	85 173

We selected the race they had run with the highest marathon correlation for each athlete, allowing us to obtain a data set with one entry per athlete. The race selected was always the marathon for athletes who had competed in the marathon. We then discarded extreme outlying data points where the predicted marathon time was greater than 11 hours, the cut-off time for the New York Marathon. This gave us our cross-race data set.

Note that while the orange points shown in figure 1A were discarded to estimate marathon times, those records were still eligible for inclusion in the cross-race data set. Fewer than 0.1% of records were discarded as outliers in creating the cross-race data set. See table 1 for a summary of the number of data points used to create our cross-race data set.

Modelling sex

NYRR does not record the sex of non-binary athletes. We addressed this by developing a probability model for the sex of each competitor. We did this by computing a variable `prob_male`, which contained our estimate of the probability that an athlete is natal-male.

For athletes who stated they were male or female, we set `prob_male` to be either 0 or 1. For non-binary athletes, we checked if we could find that athlete's sex on the Athlinks website. In this case, we set `prob_male` to be either 0 or 1.

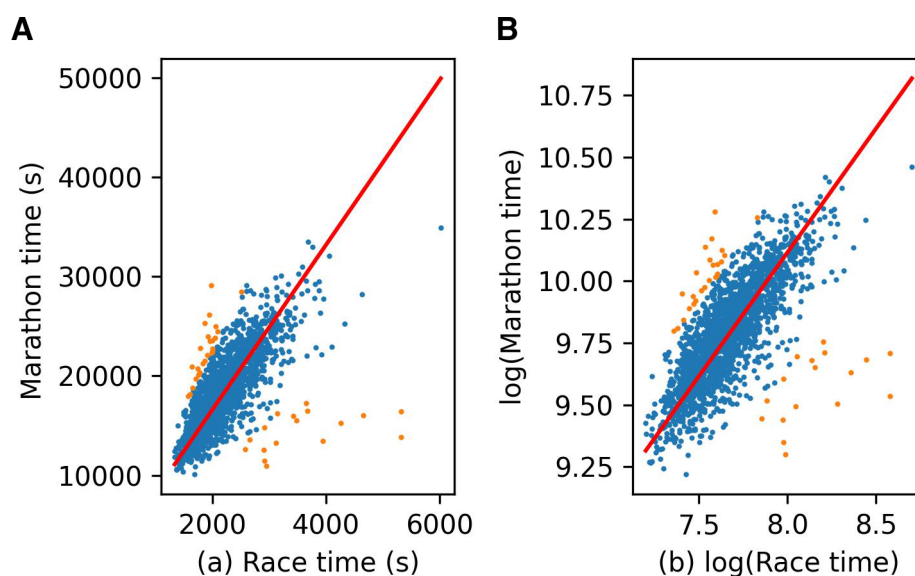


Figure 1 (A) Race time in the 'Run as One 4M' plotted against marathon time for athletes who completed both events. The slope of the red line was then estimated using points in blue. The 2% of orange points were treated as outliers, which were ignored in this computation. The red line gives the predicted marathon times for the race. (B) The same data were plotted but using a logarithmic scale.

Table 2 Summary of how we estimated the probability that a non-binary athlete was a natal-male

	No
Sex found using Athlinks	92
Probability estimated using baby name database	53
No probability assigned	21
Total	166

If we could not find the athlete on the Athlinks website, we used a database of US baby names to determine the proportion of male babies born in the US with the same first name as the athlete. We used this proportion to determine the value of prob_male. To be precise, we computed the two possible years in which the athlete was born based on their age on race day and the date of the race and used the proportion of babies born in those 2 years.

Runners where none of these methods provided any information about their sex were excluded from the analysis.

A summary of how each non-binary athlete's sex was identified is given in table 2.

We cross-validated the data from Athlinks with the probability model based on baby names. We looked at the athletes where the probability of them being a natal male was less than 0.05 or greater than 0.95 according to the baby-name database and assumed for our cross-validation that these athletes were either natal females or natal males. We found that in 77 of 78 cases where both methods allowed us to compute the athlete's sex, both methods gave the same answer.

The final cross-race data set contained records, each representing a different athlete. Each record consisted of a field event indicating which race was being run, a field gender_id indicating whether the athlete had registered as either male, female or non-binary), a field prob_male, their race time in seconds and their age.

Linear modelling

We modelled the logarithm of times rather than the times themselves. It is visually clear that the times plotted in figure 1A do not follow a multivariate normal distribution. Transforming figure 1A using a log-log plot results in the data shown in figure 1B, which is approximately normal. We see in figure 2 that modelling the logarithm of the race times rather than the race times leads to the residuals in our final model having an approximately normal distribution. It also meant we could control for the different lengths and difficulty of races simply by including an event variable in our models.

We assumed there was an unobserved variable called natal_sex, which took one of two values, 'natal male' or 'natal female'. For each runner, the probability of this variable taking the value male was assumed to be given by the variable prob_male.

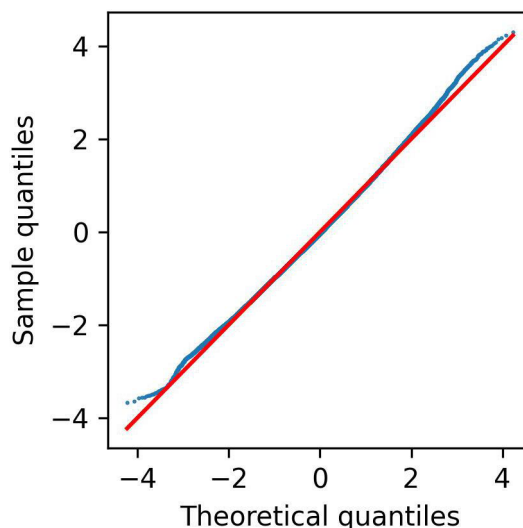


Figure 2 Normal Q-Q plot of the residuals for the model given by equation 4 for male and female athletes.

In online supplemental appendix B, we describe how to estimate the coefficients of a linear model which contains a binary unobserved variable and where some of the explanatory variables depend on this unobserved variable. The interpretation of the resulting coefficients and the p values are the same as in a conventional linear model.

Analysis

Hypothesis 1

To test the hypothesis that non-binary athletes' performance is less determined by their sex than other athletes, we created a variable nb_predictor, which took the value 1 for non-binary athletes who are natal males, -1 for non-binary athletes who are natal females and 0 otherwise. We then fit the following linear model:

$$\log(\text{time}) \sim \text{event} + \text{natal_sex} + \text{age} + \text{age}^2 + \text{nb_predictor} \quad (1)$$

The coefficients obtained for all our models can be seen in table 3. The coefficients for the intercept and the different running events are all highly significant but have been omitted as they are irrelevant to our hypotheses. We have shown p values for both one-tailed and two-tailed tests. The p values for the one-tailed tests are the ones that we use for our hypothesis testing. The two-tailed tests are the relevant values for our exploratory analyses. We have included a column showing the effect sizes as percentage increases in race times for ease of interpretation.

We have stated the coefficients of the age terms relative to 40 which is, to a good approximation, the mean age of athletes. Thus, the (age-40) coefficient represents the annual deterioration in performance at this mean age.

We used a quadratic model used for age as the data shows performance peaks at approximately age 25 then deteriorates gradually till 40 and then more sharply. With this quadratic model for age, we found no interaction terms for sex and age. We also analysed the data with a

Table 3 Coefficient estimates for each of our linear models

Parameter	Coefficient	Effect size	P value (one tailed)	P value (two tailed)
Model 1: event+natal_sex+(age-40)+(age-40) ² +nb_predictor				
sex=natal female	0.12221	13.00%***	0.0000	0.0000
(Age-40)	0.00375	0.38 %/y***	0.0000	0.0000
(Age-40) ²	0.00012	0.012 %/y ² ***	0.0000	0.0000
nb_predictor	-0.00424	-0.42 %	0.4071	0.8079
Model 2: event+gender_id+(age-40)+(age-40) ² +nb_predictor				
gender_id='female'	0.12222	13.00%***	0.0000	0.0000
gender_id='non-binary'	0.09387	9.84%***	0.0000	0.0000
(Age-40)	0.00375	0.376 %/y***	0.0000	0.0000
(Age-40) ²	0.00012	0.012 %/y ² ***	0.0000	0.0000
nb_predictor	-0.06803	-6.58%***	0.0001	0.0001
Model 3: event+gender_id+(age-40)+(age-40) ² +is_nbm+is_nbf				
sex=natal female	0.12222	13.00%***	0.0000	0.0000
(Age-40)	0.00375	0.376 %/y***	0.0000	0.0000
(Age-40) ²	0.00012	0.012 %/y ² ***	0.0000	0.0000
is_nbm	0.02584	2.62%	0.1324	0.2681
is_nbf	0.03969	4.05%	0.0580	0.1152
Model 4: event+natal_sex+(age-40)+(age-40) ² +isNB				
isNB	0.03225	3.278% (.)	0.0262	0.0528
natal_sex=natal female	0.12224	13.002 %/y***	0.0000	0.0000
(Age-40)	0.00375	0.376 %/y ² ***	0.0000	0.0000
(Age-40) ²	0.00012	0.012% ***	0.0000	0.0000

Coefficients for different events are ignored. The coefficients are given as a percentage increase in marathon time in the 'effect size' column for ease of interpretation. The final two columns contain Monte Carlo estimates for the p values of the coefficients estimated using 100 000 samples.
The symbols (.), *, **, *** indicate statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels using a two-tailed test.

more complex piecewise-linear model age, but this made no material difference to our findings.

If sex differences in race times were reduced among non-binary athletes, the sign of the coefficient for nb_predictor would be positive. However, the sign of the coefficient is negative. Thus, hypothesis 1 is not supported by our data. However, the coefficient is not statistically significant, so we cannot conclude that the sex differences in performance are greater for non-binary athletes.

Hypothesis 2

To test hypothesis 2, we fit the following linear model:

$$\log(\text{time}) \sim \text{event} + \text{gender_id} + \text{age} + \text{age}^2 + \text{nb_predictor} \quad (2)$$

In this case, the variable nb_predictor was significant at a 0.1% confidence level. Hence, we can reject the null hypothesis, and it appears that the performance of non-binary athletes is affected by their sex.

Exploratory analyses

To see if there were any other discernible associations with being non-binary, we fit the following linear model:

$$\log(\text{time}) \sim \text{event} + \text{natal_sex} + \text{age} + \text{age}^2 + \text{is_nbm} + \text{is_nbf} \quad (3)$$

Where is_nbm and is_nbf indicate if an individual is a natal male and non-binary or a natal female and non-binary, respectively, neither is_nbm nor is_nbf was significant at a 5% confidence level. The coefficients were of a very similar size. Thus, our sample size is not large enough to discern the individual effects of being natal-male and non-binary or being natal-female and non-binary, and there is no evidence of any differential effect of being non-binary between these categories. Our final choice of model is:

$$\log(\text{time}) \sim \text{event} + \text{natal_sex} + \text{age} + \text{is_nb} + \text{age}^2 \quad (4)$$

Where is_nb is a variable that indicates whether or not an athlete is non-binary.

One sees from the coefficients that being a natal female, whether or not they are non-binary, is associated with an increase in race times of approximately 13%. Being non-binary may be associated with an increase in race times of approximately 3.3% but as this is on the boundary of the 5% significance level using a two-tailed test one would wish to examine a larger sample to explore this further. The residual SD in our model for the logarithm of race times is 0.20, which corresponds to a 22% change in race times. Thus, for all athletes, the differences within sex



categories are larger than the differences between sex categories.

A q-q plot for the residuals of the model given by equation 4 for the athletes whose sex is known is shown in figure 2. This shows a good degree of normality for the residuals, validating this assumption of our hypothesis testing.

DISCUSSION

Our results illustrate the value of data on sex and gender identity.

The differential between natal male and natal female performances is better explained by differences in sex than differences in gender identity, as this differential persists for our non-binary cohort. This provides evidence against the theory that an individual's gender-identity plays a significant role in these disparities in addition to their sex.

Our exploratory analysis indicates that non-binary athletes may have slower race times than other athletes once one controls for sex and age, but one would wish to confirm this with a larger data set as this is on the boundary of statistical significance. Data gathered on gender non-conforming college students by the American College Health Association³⁶ suggest that gender non-conforming students are less likely to meet exercise recommendations, have increased rates of obesity and have higher rates of physical and mental health issues; these factors affect levels of fitness and training status. A complex range of factors associated with non-binary status could account for any association with slower race times. We do not wish to suggest causality in either direction.

Research implications

Any possible differential in the performance of non-binary athletes would be masked if one did not consider sex. This illustrates that if one wishes to understand the needs of gender non-conforming individuals, it is vital to control for sex as it is likely to play a significant role in any analysis.

The prediction arising from gender-identity theory that sex differences in performance will be lower among non-binary athletes in mass-participation running events is not supported by our results. Our study illustrates the importance of controlling for sex when studying gender non-conforming individuals in the context of sports performance. When considering gender-identity theory in any context, one cannot rely purely on theoretical arguments to determine whether gender-identity or sex is the more significant factor.

Given the lack of empirical evidence supporting gender-identity theory, one should not assume by default that gender-identity is a more powerful explanatory variable than sex. Being an objectively measurable binary variable, sex has considerable explanatory advantages over gender identity.

Our study illustrates that gender identity may be significant as a variable even in situations where the data do not support the predictions of gender-identity theory. It would be interesting to see in a larger study whether non-binary athletes do have slower race times controlling for sex and age, if so this would justify our initial assumption that social factors have a significant impact on sporting performance.

In summary, our study shows the value of gathering data on sex and gender-identity. It illustrates how this is particularly necessary if one wishes to understand the experiences of gender non-conforming individuals. A similar view is put forward in Hunter *et al.*⁹

Limitations

Non-binary runners may have chosen to run as either male or female athletes. Thus, our analysis only applies to those non-binary athletes who chose to run in the non-binary category. The race categories conflated sex categories (male and female) with gender identity categories such as non-binary. Our sample of non-binary athletes is not much larger than needed to detect sex differences, so if gender identity differences exist but are smaller than sex differences our study may not have had sufficient power to detect them. Our analysis has grouped all non-binary identities into a single category, if sufficient data and suitable categories were available, a more nuanced analysis of non-binary identities might reveal effects for specific forms of non-binary identity.

Contributors All authors listed met the conditions required for full authorship. JA and AS performed the analysis. GMP collected the data. JA drafted the manuscript which was amended by AS and GMP and approved by all for submission. JA is responsible for the overall content as guarantor.

Funding AS is a member of the Centre for Longitudinal Studies, an Economic and Social Research Council resource centre funded by ES/M001660/1.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants but not applicable. Under King's College London ethics policy, ethical approval was not required for this study as it was a secondary data analysis using data already in the public domain. This study is an analysis of data already in the public domain. Participants have consented to make their data publicly available.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given,

and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iD

John Armstrong <http://orcid.org/0000-0002-4232-9555>

REFERENCES

- Pike J. Safety, fairness, and inclusion: transgender athletes and the essence of rugby. *J Philos Sport* 2021;48:155–68.
- Burke M. Trans women participation in sport: a feminist alternative to pike's position. *J Philos Sport* 2022;49:212–29.
- Devine C. Female olympians' voices: female sports categories and international olympic committee transgender guidelines. *Int Rev Sport Sociol* 2022;57:335–61.
- Kroshus E, Ackerman KE, Brown M, et al. Improving inclusion and well-being of trans and gender nonconforming collegiate student-athletes: foundational concepts from the national collegiate athletic association summit on gender identity and student-athlete participation. *Br J Sports Med* 2023;57:564–70.
- Martowicz M, Budgett R, Pape M, et al. Position statement: IOC framework on fairness, inclusion and non-discrimination on the basis of gender identity and sex variations. *Br J Sports Med* 2023;57:26–32.
- Erikainen S, Vincent B, Hopkins A. Specific detriment: barriers and opportunities for non-binary inclusive sports in Scotland. *J Sport Soc* 2022;46:75–102.
- Sullivan A, Todd S. Two sexes. In: *Sex and gender*. Routledge, 2023: 16–34.
- Dawkins R. Why biological sex matters; 2023. New statesman (London, England: 1957)
- Hunter SK, Angadi SS, Bhargava A, et al. The biological basis of sex differences in athletic performance: consensus statement for the American college of sports medicine. *Transl J ACSM* 2023;8:1–33.
- Butler J. Gender trouble. routledge;
- Stonewall. List of LGBTQ+ terms. 2023. Available: <https://www.stonewall.org.uk/list-lgbtq-terms>
- Stonewall. The truth about Trans. 2023. Available: <https://web.archive.org/web/20230818081341/https://www.stonewall.org.uk/the-truth-about-trans>
- Safer JD. A current model of sex including all biological components of sexual reproduction. *Law & Contemp Probs* 2022;85:47.
- Sullivan A. Sex and the office for national statistics: a case study in policy capture. *Political Quarterly* 2021;92:638–51.
- Sullivan A, Murray K, Mackenzie L. Why do we need data on sex? In: Sullivan A, Todd S, eds. *Sex and gender: A contemporary reader*. Routledge, 2023.
- Sullivan A. Sex and the census: why surveys should not conflate sex and gender identity. *Int J Soc Res Methodol* 2020;23:517–24.
- Fugard A. Should trans people be postmodernist in the streets but positivist in the spreadsheets? A reply to sullivan. *Int J Soc Res Methodol* 2020;23:525–31.
- Halliday R. Sex, gender identity, Trans status - data collection and publication: guidance; 2021. Scottish government
- Dhejne C, Lichtenstein P, Boman M, et al. Long-term follow-up of transsexual persons undergoing sex reassignment surgery: cohort study in Sweden. *PLoS One* 2011;6.
- Hilton EN, Lundberg TR. Correction to: transgender women in the female category of sport: perspectives on testosterone suppression and performance advantage. *Sports Med* 2021;51:2235.
- Harper J, O'Donnell E, Sorouri Khorashad B, et al. How does hormone transition in transgender women change body composition, muscle strength and haemoglobin? systematic review with a focus on the implications for sport participation. *Br J Sports Med* 2021;55:865–72.
- Briane P. *Employment Tribunal rulings on gender-critical beliefs in the workplace*. House of Commons Library, 2022.
- Scruton S, Flintoff A. *Gender and sport: A reader*. Psychology Press, 2002.
- Theberge N. Gender and sport; 2000. Handbook of sports studies 322–33.
- Hines S. *TransForming gender: Transgender practices of identity, intimacy and care*. Policy Press, 2007.
- Richards C, Bouman WP, Seal L, et al. Non-binary or genderqueer genders. *Int Rev Psychiatry* 2016;28:95–102.
- Human Rights Campaign. Glossary of terms. 2023. Available: <https://web.archive.org/web/20230928150543/https://www.hrc.org/resources/glossary-of-terms>
- Barras A, Frith H, Jarvis N, et al. Timelines and transitions: understanding Transgender and non-binary people's participation in everyday sport and physical exercise through a temporal lens. In: *Temporality in qualitative inquiry theories, methods and practices*. 2021: 57–71.
- Caudwell J. Transgender and non-binary swimming in the UK: indoor public pool spaces and UN/safety. *Front Sociol* 2020;5:64:64..
- Spandler H, Erikainen S, Hopkins A, et al. Non-binary inclusion in sport: Uclan publishing. 2020.
- Whitehouse L, Armstrong C, Cooke J. The navigation of non-binary expression through sports kit; 2022. Gender diversity and sport: Routledge 77–96.
- van Buuren S. Flexible imputation of missing data. 2nd edn. CRC press, 2018.
- NYRR race results. New York Road Runners; 2023. Available: <https://results.nyrr.org/home>
- Athlinks. Official race results finder. 2023. Available: <https://blog.athlinks.com/results>
- US Social Security Administration. Baby name database. 2023. Available: <http://www.ssa.gov/OACT/babynames/names.zip>
- American College Health Association. NCHA III national college health assessment; 2021.

Appendix A

The races included in our data set were:

Race New York Marathon
NYRR Staten Island Half
New Balance Bronx 10M
RBC Brooklyn Half
NYRR Team Championships 5M
NYRR Grete's Great Gallop 10K
Mastercard New York Mini 10K
NYRR Queens 10K
Fred Lebow Half-Marathon
NYRR Ted Corbitt 15K
Percy Sutton Harlem 5K
NYRR Washington Heights Salsa, Blues, and Shamrocks 5K
Italy Run by Ferrero 4M
Gridiron 4M presented by The FLAG Art Foundation
NYRR Joe Kleinerman 10K
NYRR Manhattan 10K
Front Runners New York LGBT Pride Run 4M
New Balance 5th Avenue Mile
Run as One 4M presented by JPMorgan Chase
Achilles Hope & Possibility 4M; NYRR Al Gordon 4M
NYRR Newport 5K

We excluded the "SHAPE + Health Women's Half Marathon" as it was a women-only event and the "TCS New York City Marathon Training Series 12M" as it was a training event rather than a race.

Appendix B – Linear modelling with a hidden variable

Suppose that from our sample we can compute the values of d_1 exogenous variables in X^1, X^2, \dots, X^{d_1} . Suppose that there are d_2 additional variables Y^1, Y^2, \dots, Y^{d_2} which we cannot compute directly from our sample because they depend upon a hidden Bernoulli variable J , but we do know that we compute the values T^1, T^2, \dots, T^{d_2} that they will take if the Bernoulli variable is equal to 0 and we can also compute the values F^1, F^2, \dots, F^{d_2} that they will take if the hidden Bernoulli variable is equal to 1. We also have an exogenous variable p that contains the probability with which the value 1 occurs.

In our applications, J takes the value 1 if the athlete is female and 0 if the athlete is male. The variable Y^1 might then represent the sex of the athlete and may also take the value 1 if the athlete is female and 0 otherwise. In this case $T^1 = 1$ and $F^1 = 0$. The variable Y^2 might be used to represent the interaction between sex and age, so we would set T^2 to equal the variable **age** and F^2 to equal 0.

We would like to model an endogenous variable Z using a linear model in our exogenous variables of the form:

$$Z = \sum_{i=1}^{d_1} \beta_i X^i + \sum_{i=1}^{d_2} \gamma_i Y^i + \varepsilon$$

where the ε are assumed to be independent identically distributed random variables of finite variance and mean 0. We assume these variables are also independent of J . This is the form of model used when we describe our analysis in this paper. When we come to compute P-values, we will further assume that the ε follow a normal distribution.

Since $Y^i = J(T^i - F^i) + F^i$ we may rewrite this equation as follows:

$$Z = \sum_{i=1}^{d_1} \beta_i X^i + \sum_{i=1}^{d_2} \gamma_i (p(T^i - F^i) + F^i) + \varepsilon + \sum_{i=1}^{d_2} \gamma_i (J(T^i - F^i) - p(T^i - F^i))$$

Equation 1

If we define a new exogenous variable

$$P^i := p(T^i - F^i) + F^i$$

and a new hidden variable

$$\tilde{\varepsilon} := \varepsilon + \sum_{i=1}^{d_2} \gamma_i (J(T^i - F^i) - p(T^i - F^i))$$

then our model becomes a more familiar linear model where all terms are known in the sample apart from the single random term $\tilde{\varepsilon}$.

$$Z = \sum_{i=1}^{d_1} \beta_i X^i + \sum_{i=1}^{d_2} \gamma_i P^i + \tilde{\varepsilon}.$$

Equation 2

Since the expectation of J is equal p , we see that expectation of $\tilde{\varepsilon}$ is zero. This means we may use ordinary least squares to estimate the coefficients in this regression. However, the $\tilde{\varepsilon}$ will no longer be identically distributed: instead, they will have variance given by

$$V := \text{Var } \tilde{\varepsilon} = \left(\sum_{i=1}^{d_2} \gamma_i (T^i - F^i)^2 \right)^2 p(1 - p) + \text{Var } \varepsilon.$$

Equation 3

As a result, the ordinary least squares estimator for Equation 2 will not be an optimal estimator in the sense of the Gauss-Markov theorem. To obtain a better estimation of our coefficients we adopt a two-stage estimation process which we will now describe.

First, we use ordinary least squares to get a first estimate of the γ_i . Next, we estimate $\text{Var } \varepsilon$ using the equation:

$$N \text{ Var } \varepsilon + \sum_{\alpha=1}^N \left(\sum_{i=1}^{d_2} \gamma_i (T^i - F^i)^2 \right)^2 p(1-p) = \sum_{\alpha=1}^N V_{\alpha} \approx \frac{N}{N - d_1 - d_2} \sum_{\alpha=1}^N r_{\alpha}^2.$$

Equation 4

In this equation, N denotes the sample size, the index α runs over the items in the sample and r_{α} denotes the residual for item α . This equation is derived from Equation 3 combined with the standard estimate for the total variance of the model. Using Equation 3 we now know the variance of the ε .

The second step of our estimation process is to re-estimate the coefficients of our model using weighted least squares with weights $\frac{1}{V_{\alpha}}$. By the Gauss-Markov theorem this should approximate the best linear approximator. We can also compute a new estimate for improved estimator for $\text{Var } \varepsilon$ using a weighted version of Equation 4:

$$\left(\text{Var } \varepsilon \sum_{\alpha=1}^N \frac{1}{V_{\alpha}} \right) + \sum_{\alpha=1}^N \frac{1}{V_{\alpha}} \left(\sum_{i=1}^{d_2} \gamma_i (T^i - F^i)^2 \right)^2 p(1-p) \approx \frac{N}{N - d_1 - d_2} \sum_{\alpha=1}^N \frac{1}{V_{\alpha}} r_{\alpha}^2$$

To compute the P-values of the coefficients, we used Monte Carlo simulations. We then computed the P-values using a 2-tailed test with a Monte Carlo simulation assuming that the ε are normally distributed. In detail, we simulated J and ε on the basis of this assumption, and so could compute Z using Equation 1 for any desired choice of coefficients β and γ . To compute the P-value of a particular coefficient, we performed Monte Carlo simulations under the null hypothesis that the coefficient was instead 0. We then counted the proportion of occasions on which our estimation procedure gave a larger coefficient value than the parameter estimate arising from the data.

Appendix C – Modelling additional uncertainty

We have described in the paper how we computed a variable called **prob_male** which we use to model the probability that an athlete is natal-male based on their name and race category. There is no “correct” model for this probability and other models can be proposed. In particular, since athletes may have changed their name since birth, one might want to add in some additional uncertainty to the model to reflect that. A simple way to do this is to choose a parameter value α which takes values between 0 and 0.5 and to then define a new variable **q** as follows:

$$\mathbf{q} = \begin{cases} \mathbf{prob_male}, & \text{athlete is not non-binary} \\ \alpha + (1 - 2\alpha)\mathbf{prob_male}, & \text{athlete is non-binary.} \end{cases}$$

The variable **q** models the probability that an athlete is male, with some additional uncertainty added over and above that arising from the distribution of given names. If a non-binary athlete has a name which is only used for females, then the **q**-model probability of them being a natal male will be α . If they have a name which is only used for males, then the **q**-model probability of being a natal male will be $1 - \alpha$. Hence α is a measure of the uncertainty in the **q**-model on top of the uncertainty arising from their name alone.

If we re-run our analysis using **q** in place of **prob_male**, we can test our hypotheses using this alternative probability model. The results are shown in Table 5 in the case when $\alpha = 0.05$. Given how successful our cross-validation was at predicting the natal sex of non-binary athletes, we feel

this choice of α is probably an over-estimate of the uncertainty arising from issues such as changes of name.

Parameter	Coefficient	Effect size	P value (1 tail)	P value (2 tail)
Model 1: event + natal_sex + (Age-40) + (Age-40)² + nbPredictor				
natal_sex='natal_female'	0.12221	13.0 % ***	0.0000	0.0000
(Age-40)	0.00375	0.38 %/y ***	0.0000	0.0000
(Age-40) ²	0.00012	0.012 %/y ² ***	0.0000	0.0000
nbPredictor	-0.06957	-6.7 % *	0.0227	0.0466
Model 2: event + gender + (Age-40) + (Age-40)² + nbPredictor				
gender='female'	0.12221	13.0 % ***	0.0000	0.0000
gender='non-binary'	0.09388	9.84 % ***	0.0000	0.0000
(Age-40)	0.00375	0.38 %/y ***	0.0000	0.0000
(Age-40) ²	0.00012	0.012 %/y ² ***	0.0000	0.0000
nbPredictor	-0.13608	-12.723 % ***	0.0001	0.0001
Model 3: event + natal_sex + (Age-40) + (Age-40)² + is_nbm + is_nbf				
natal_sex='natal female'	0.12221	13.0 % ***	0.0000	0.0000
(Age -40)	0.00375	0.38 %/y ***	0.0000	0.0000
(Age -40) ²	0.00012	0.012 %/y ² ***	0.0000	0.0000
is_nbm	-0.04220	-4.13 %	0.1320	0.2617
is_nbf	0.10775	11.4 % **	0.0035	0.0075
Model 4: event + natal_sex + (Age-40) + (Age-40)² + isNB				
isNB	0.02994	3.0 %	0.0363	0.0727
natal_sex='natal female'	0.12227	13.0 % ***	0.0000	0.0000
(Age -40)	0.00375	0.376 %/y ***	0.0000	0.0000
(Age -40) ²	0.00012	0.012 %/y ² ***	0.0000	0.0000

Table 5: Coefficient estimates for each of our linear models using the q -model when $\alpha=0.05$. The final two columns contain Monte Carlo estimates for the P-values of the coefficients estimated using 100,000 samples. The asterisks indicate statistical significance at the 0.05, 0.01 and 0.001 levels using a 2-tailed test.

Using the **q**-model with $\alpha = 0.05$ we find that there is a statistically significant widening of the sex gap in athlete's performance for non-binary runners. When we choose an implausibly high value for the uncertainty α this effect is more exaggerated. This property of our model is easily explained: the range of the predictor variable (the probabilities of being male or female) is smaller in the **q**-model than in the original model, but the range of the outcome variables (race times) are unchanged. As a result, one should expect a larger magnitude for the coefficient of **nb_predictor** in the **q**-model and one expects its magnitude to increase as α increases. This shows that our original approach of ignoring the uncertainty arising from athlete's changing may lead us to slightly under-estimate the magnitude of sex differences in running performance among non-binary athletes, but assuming that $\alpha < 0.05$, any bias introduced this way will be a relatively small.